
The Need for Speed

Accelerated Computing for Financial Applications

The need for speed

This article outlines the findings of a review of accelerated computing options carried out on behalf of one of our investment banking clients. There is a crisis emerging in computational power in The City: over the last few years the demand for higher processing capacity for financial systems has increased rapidly. This has been driven by a combination of higher trading volumes, increasing exposures resulting in additional risk calculations, greater intra-day requirements and higher throughput demands for semi- and fully-automated electronic market trading systems. Current computing technology has kept pace but is beginning to reach its limits. As CPUs become more powerful in GFlops terms the demands for power and cooling rise at an exponential rate. Physical space in data centres is finite and the laws of physics and economics start limiting the computation power available. Consequently, some new approaches need to be considered to gain more speed within these same constraints.

Faster but not at any cost

One critical consideration of the review was the non-functional characteristics of any solution, including price. Increasing calculation capacity is the primary goal but a faster performing solution cannot come at the expense of the ability to construct and maintain the solution. Non-functional factors considerably impact the total cost of a solution: the thermal impact (i.e. how much the air conditioning bill will increase/decrease if the solution is deployed); integration of the solution into the existing infrastructure; and the need to change development process or introduce new technical skills.

Won't CPUs save the day?

We are used to the near-linear increase in CPU speed: every eighteen months a new CPU is released that is roughly twice as fast as the previous generation. The hidden story of this performance increase is the exponential increase in heat output. Over the last 10 years server chip performance has increased by about 40x whereas heat output has increased 60x¹. Manufacturers are struggling with the physical implications of this and their customers are struggling with the financial implications: 70% of data centre costs are due to power consumption and cooling requirements².

The introduction of multi-core CPUs is partially motivated by the need to rein in the impact of heat output. For each chip a headline 'equivalent' CPU speed is quoted but the cores are each clocked at a slower speed – hence reducing

¹ Using SpecFP2000 figures, die size and heat density data information from Intel

² Malone, C., Belady, C. "Metrics to characterize Data Center & IT Equipment Energy Use," Proceedings of 2006 Digital Power Forum, Richardson, Texas, September 2006

heat output. 'Equivalent' speed is achieved in multi-threaded applications where the operating system can distribute threads across the cores and so the overall throughput is comparable to a CPU of the quoted speed. The story is very different for single-threaded applications, however: *a single threaded application will run more slowly on a multi-core CPU than on a single core CPU of the equivalent speed.*

Multi-core chips do provide a path to better computational throughput provided appropriate multi-thread applications are written. However, to achieve greater numbers of cores on the chips manufactures are going to have to decrease gates size (the physical size of transistors) or increase the die size (the physical size of the chip) both of which will result in increases in heat output that will rise exponentially with performance gain.

Accelerated computing

In our consideration of alternative accelerated computing solutions we considered two main types of technology: FPGAs and array processors. Other technologies, such as hybrid supercomputers consisting of arrays of CPUs and FPGAs, were ruled out due to issues with cost and availability. In both cases the solution introduces a co-processor to work in concert with the CPU, with the co-processor outputting significantly less heat than the CPU.

FPGAs

FPGA chips are packed with the same basic low-level logic building blocks as those used in general-purpose CPUs and application-specific chips such as analogue-digital converters. The big difference is that the logic blocks are 'un-configured' and a hardware design can be 'flashed' onto the chip to define the connections between them. An existing design can be erased and a new design flashed, allowing the same chip to be reconfigured for many different purposes.

The promise of FPGAs is that the implementation of an algorithm or part of an algorithm can be optimised beyond the constraints of a general-purpose CPU approach. Various techniques can be used to undertake calculations at a much higher rate than could be achieved on a CPU. Typically FPGAs run at much lower clock speeds than CPUs, thus producing much less heat.

Array Processors

Array processors (also known as SIMD processors) work on the principle of executing the same instruction across multiple sets of data simultaneously. Data is split into arrays with each processing element in the array processor acting in the same way on a different data element in the data array. As with FPGAs, the clock speed of the array processor is lower than a CPU. However, producing an array of results on each clock cycle rather than a single result means that results are effectively produced at an accelerated rate.

FPGA vs SIMD

In order to compare the performance and non-functional characteristics of these two very different acceleration approaches we implemented accelerated versions of both a standard Black-Scholes Euro Option Pricer and a Collateralized Debt Obligation pricing algorithm provided by our customer. For both we required that the solution used double-precision floating point representation and the coding of the algorithm was comprehensible by software developers. This latter point is of particular relevance to FPGA-based solutions as the semantics of some of the low-level hardware design languages are not easily understood by software developers and the algorithm is obscured by layout and routing information. This comprehension is required for both maintainability and productivity. Both solutions had very similar heat output characteristics and similar cost.

For our FPGA-based solution we selected an offering from Celoxica, a company that packages third-party FPGA boards with tools that allow programs to be written in their own proprietary C-like language and then 'synthesised' to a hardware design, plus a range of libraries including one for arbitrary position maths.

For our SIMD solution we selected an offering from Clearspeed, a company that sells their own double-precision array processor, with 96 processing elements per chip, along with a proprietary C-like language.

Comparison results

Aspect	FPGA	SIMD
Actual performance acceleration ³	1-2x. Both Euro Option and CDO.	2-92x. 2x for basic CDO port, 10x for optimised CDO port, 92x for optimised Euro Option port.
Limitations on acceleration	<p>Size of FPGA. Only a very small portion of both Euro Option and CDO algorithms would fit on our chip due to the number of logic blocks necessary to implement double-precision maths.</p> <p>Throughput of interconnect speed. With only a small portion of the algorithm on the chip, communication between CPU and FPGA co-processor becomes the limiting factor.</p>	<p>Degree of branching in the algorithm. The Euro Option has no branching so acceleration is directly proportional to the number of processing elements in the array. The CDO has a number of branch and convergence points and so the acceleration was much lower than that for the Euro Option.</p>
Best strategy for acceleration	<p>Highly optimised implementation of small bottleneck(s) in the algorithm. Best strategy for acceleration was implementation a highly-optimised pseudo-random number generator.</p>	<p>Complete port of algorithm to array-based calculation. Both Euro Option and CDO ran in their entirety on the SIMD solution.</p>

³ In comparison to 3.0Ghz Intel Xeon executing equivalent ANSI C code. Range describes findings across a variety of approaches and implementation approaches.

<p>Impact on development approach</p>	<p>'C-like' language still requires hardware design skills. Whilst it looks a bit like C, the hardware design roots of the proprietary language were clearly evident and hardware design considerations are never far from the developers mind.</p> <p>Unproductive tools. Compile/synthesize times of 2 hours to 3 days means a top-down design-then-implement approach is forced on the developers. Poor debugging tools compound this problem: if the solution isn't producing the right answers, debugging and correcting the defect will be very unproductive.</p>	<p>Relatively small shift in development paradigm. Programming for single instruction multiple data does require a shift in development paradigm away from standard single instruction sign data but it is easy for competent programmers to adopt.</p> <p>Opportunity for incremental development approach. The SIMD solution will run an unmodified ANSI C program on a single processing element. Parts of the algorithm can be refactored to an SIMD approach in an incremental manner; maintaining a working solution throughout the optimisation process, ensuring alteration brings improvement, and allowing a balance of effort vs. improvement.</p>
<p>Suitability</p>	<p>Small-scale, highly optimizable, low-data, algorithms such as pseudo-random number generator or simple option pricer.</p>	<p>Large-scale, branch-free, medium data algorithms such as CDO or BGM.</p>

Conclusions

Although FPGAs promise much they fail to deliver within the context of the parameters we set for the review. We do know of financial applications that successfully use FPGAs for acceleration of pricing and valuation calculations but these have relied on direct hardware design rather than use of a high-level 'programming' language, compromised on calculation precision using fixed-point maths rather than floating point, and tolerate lengthy development and modification time-scales. Similar or better results could be achieved with the Celoxica solution but the issues described above stem from the current state of FPGA technology rather than the specific tool reviewed. As both the hardware and supporting tools improve, these restrictions will relax and FPGAs will become a more realistic proposition. We expect some significant improvements over the next 24 months.

The SIMD solution has much greater potential to live up to its promise at the present time. Depending on the degree of branching in the algorithm (and, to a much lesser extent, the size of the dataset) acceleration is significant for relatively little effort; the SIMD CDO was ported and optimised from the ANSI C version over a period of three days. Of course, all institutions will be optimising their pricing code far beyond an the limitation of ANSI C and so an optimised SIMD solution will not offer 92x improvements over a single CPU solution with the same optimisation effort applied. However, if acceleration per dollar spent on both development and hosting is considered, we believe SIMD co-processors can offer significant benefits.

About e2x

e2x helps its clients achieve better, faster, smarter software delivery through a mixture of consultancy and project delivery. In the engagement described, e2x brought its significant experience of financial systems, advanced technology and sustainable development and delivery approaches to bear on a problem that encompasses all three areas.

The authors

John S. Nolan is a partner in e2x. John has many years of experience working in a number of domains from finance and advertising through to engineering and research. John is also a certified ScrumMaster, speaker at international conferences and an ACM Distinguished Engineer.

Paul Dyson is a partner in e2x. Paul is a practising advocate and early pioneer of agile development methodologies and has applied agile approaches to many domains from investment banking to retail. Paul is a published author and speaker at a number of international conferences.

Contact us

e2x limited
35 New Broad St.
London EC2M 1NH

t: +44 (0)207 194 8015

f: +44 (0)207 194 8016

www.e2x.co.uk

john@e2x.co.uk

paul@e2x.co.uk